INFO 6210
Data Management and Database Design

Gathering, Scraping, Munging and Cleaning Data

Assignment 1- Crime Database

Team:

Ashwin Lakshman
001353233

Navaneeta Naik
001027107

CONTENTS

ABSTRACT

Data munging/wrangling process is performed on real world data where the database tables are populated with it.

The Denver crimes statistical data is chosen for data munging and analysis purposes. This dataset includes criminal offenses in the City and County of Denver. Thus, we can statically analyze the wide variety of details used to describe the crime, to find relationships between them and to cluster the crimes according to some criteria. In the rest of the report we will explore the crimes and their corresponding variables.

Data Sources

Data having thematic relationship is gathered from 3 different sources.

1. A web scraper
2. A web API
3. CSV format Dataset

Web scraper

Web Scraping, also termed as Screen Scraping, Web Data Extraction, Web Harvesting etc. is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format, for later retrieval or analysis. Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining.

Denver crime's statistical data that includes offense code, offense type, offense category, is_crime, is_traffic is scraped from below website. The data received in the form of
JSON is converted into DataFrame.

Since the webpage Kaggle.com makes use of AJAX to retrieve tables, it cannot be parsed using BeautifulSoup. Hence, we created a webpage using the table and parsed it as a HTML file.

Website: https://www.kaggle.com/paultimothymooney/denver-crime-data

DataFrame is stored in the variable new_df.

Web API

A Web API is an application programming interface for either a web server or a web browser.

Advantages of using Web API are:

- It retrieves data in bulk or with great specificity which would be time consuming otherwise.
- It provides a way to access information that doesn't exist on the web and are only stored in a database attic which is hidden from users.
- It automates a news app that needs live data from other sources.
- It provides more direct interface for reading and writing data to a service.

CSV format Dataset

A common text-based data interchange format is the comma-separated value (CSV) file. This is often used when transferring spreadsheets or other tabular data.

Kaggle Dataset is used as the third resource where data is gathered in the form of CSV file. Kaggle is a platform for predictive modelling and analytics competitions in which companies and researchers post data and statisticians and data miners compete to produce the best models for predicting and describing the data.

Denver crime's statistical data that includes offense code, offense type, offense category, is_crime, is_traffic are present in the CSV file extracted from below website:

https://www.kaggle.com/paultimothymooney/denver-crime-data

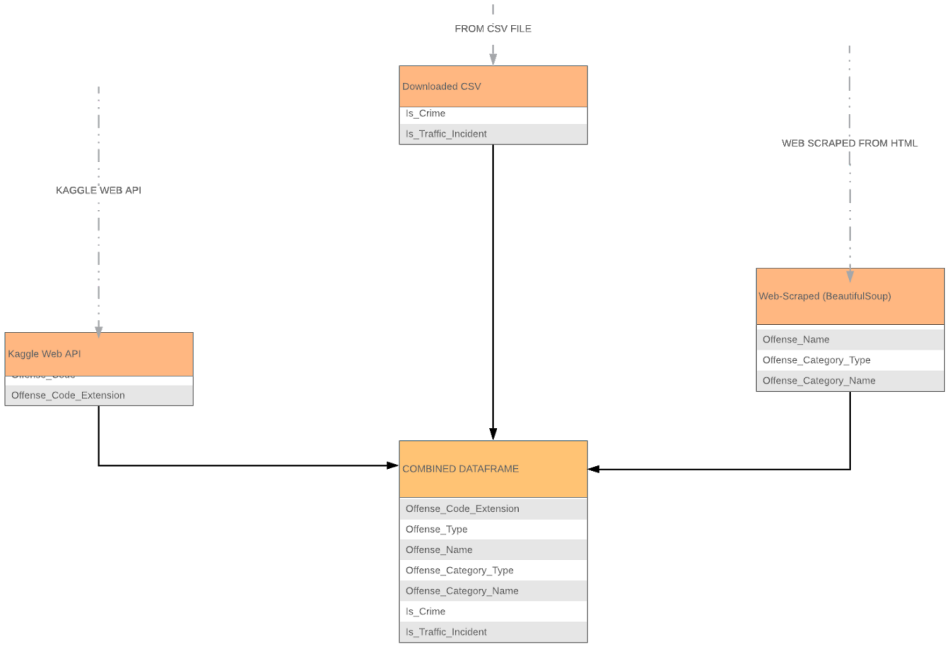
Data Fields/Variables

- **Offense_code:** Code of the offense
- **Offense_type:** Type of the offense
- **Cat_name:** Name of the category of the crimes
- **Cat_type :** Type of the category of the crimes
- **Is_crime:** If it's a crime and not a traffic incident
- **Is_traffic:** If it's a traffic incident and not a crime

Conceptual Database Model

A conceptual schema or conceptual data model is a map of concepts and their relationships used for databases. This describes the semantics of an organization and represents a series of assertions about its nature. Specifically, it describes the things of significance to an organization (entity classes), about which it is inclined to collect information, and characteristics of (attributes) and associations between pairs of those things of significance (relationships).

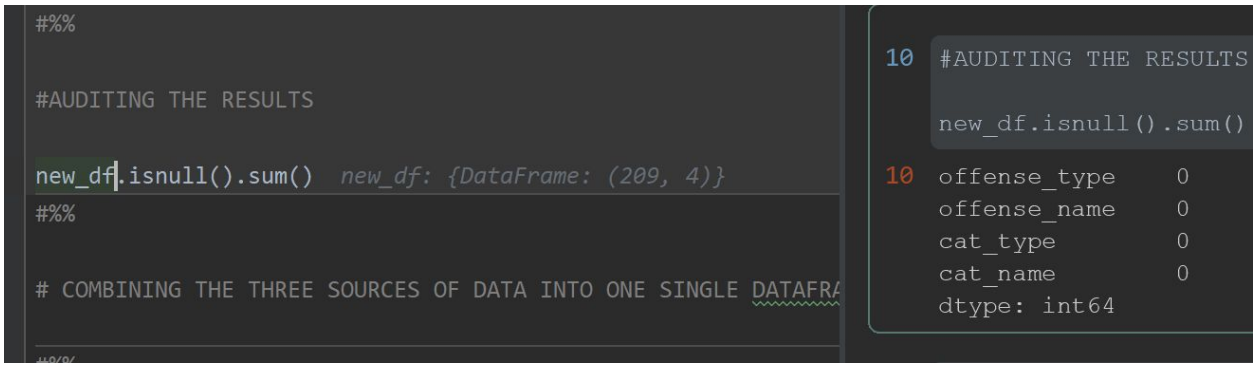
ER DIAGRAM


FROM CSV FILE
Downloaded CSV
Is_Crime
Is_Traffic_Incident
KAGGLE WEB API
Kaggle Web API
Offense_Code_Extension
WEB SCRAPED FROM HTML
Web-Scraped (BeautifulSoup)
Offense_Name
Offense_Category_Type
Offense_Category_Name
COMBINED DATAFRAME
Offense_Code_Extension
Offense_Type
Offense_Name
Offense_Category_Type
Offense_Category_Name
Is_Crime
Is_Traffic_Incident

Data Auditing

Data gathered from all three sources are validated by detecting and correcting inaccurate records. The modified data is then checked for consistency and displayed in the form of dataframe.

Data cleaning process is performed on web scraped data as follows:



Data cleaning process is performed on data obtained through web API as follows:

Data cleaning process is performed on data obtained through Kaggle CSV is as follows:

Conclusions

Data gathered from web scraping process.

```
int("HIHI")
rint(offense)
0
n1=len(offense)-3  offense: ['stolen-property-possessio
ile i<=len1:  i: 836  len1: 834
  offense_type.append(offense[i])  offense_type: ['stol

  offense_name.append(offense[i+1])  offense_name: ['Po

  cat_type.append(offense[i+2])  cat_type: ['all-other-

  cat_name.append(offense[i+3])  cat_name: ['All Other
  i+=4  i: 836

reating dataframe and storing the scrapped data

.set_option('display.max_colwidth', -1)
w_df = pd.DataFrame({"offense_type":offense_type,"offen
```

	offense_type	offense_name	cat_type	cat_name
0	stolen-property-possession	Possession of stolen property	all-other-crimes	All Other Crimes
1	fraud-possess-financial-device	Possession of a financial device	all-other-crimes	All Other Crimes
2	damaged-prop-bus	Damaged business property	criminal-mischief-private	Criminal mischief to private property
3	criminal-mischief-public	Criminal mischief to public property	criminal-mischief-other	Criminal mischief - other
4	criminal-mischief-mtr-veh	Criminal mischief to a motor vehicle	criminal-mischief-graffiti	Criminal mischief - graffiti

Data gathered from Web API.

```
K 2 : MANIPULATE DATA USING WEB API


= kaggle.KaggleApi()
authenticate()  akg: <kaggle.api.kaggle_api_extended.
asetloc = 'paultimothymooney/denver-crime-data'
aset = akg.dataset_view(adatasetloc)  akg: <kaggle.ap
aset.files  adataset: paultimothymooney/denver-crime-

dataset_download_cli(adatasetloc, unzip=True, force=T
```

```
12 newds
```

12

	off_code	offcode_ext
0	2804	1
1	2804	2
2	2901	0
3	2902	0
4	2903	0
...	...	...

Data gathered from CSV format dataset.

```
#TASK 1:   READING DOWNLOADED CSV FILE AND STORING AS A DATAFRAME.
import kaggle
import pandas as pd
import re
offense_description = pd.read_csv("C:/Users/ashwi/Downloads/denver-crime-data/offense_codes.csv",enco
extracted= offense_description[['IS_CRIME','IS_TRAFFIC']].copy()
extracted
```

	IS_CRIME	IS_TRAFFIC
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0
...	...	...

Combined Data

The data from the 3 sources were combined to form a single collaborative dataframe.

	off_code	offcode_ext	offense_type	offense_name	cat_type	cat_name	IS_CRIME	IS_TRAFFIC
0	2804	1	stolen-property-possession	Possession of stolen property	all-other-crimes	All Other Crimes	1	0
1	2804	2	fraud-possess-financial-device	Possession of a financial device	all-other-crimes	All Other Crimes	1	0
2	2901	0	damaged-prop-bus	Damaged business property	criminal-mischief-private	Criminal mischief to private property	1	0
3	2902	0	criminal-mischief-public	Criminal mischief to public	criminal-mischief-other	Criminal mischief - other	1	0

References

https://www.kaggle.com/paultimothymooney/denver-crime-data